新聞記事における偏向性の定量評価

新聞記事における偏向性の定量評価

QUANTITATIVE EVALUATION OF BIAS AMONG NEWSPAPER ARTICLES

小谷　龍ノ介
Ryunosuke KOTANI
指導教員　彌冨　仁

法政大学大学院理工学研究科応用情報工学専攻修士課程

It is necessary to select from a vast amount of information in order to obtain highly reliable information from the web media. Considering the objectivity, accuracy, fairness, etc. of the information to be selected, understanding of the bias of the media is required. In various media, it is difficult to understand at a glance how difference the bias is, and it is difficult to receive comprehensively and accurately and fairly information. In this study, focusing on Web news, we aim to quantitatively visualize the bias of each newspaper company. In order to visualize the bias, "characteristic words" which can interpret the sentiment are extracted, and the way of handling each newspaper company for them is quantified. As the quantitative index, we used the sentiment value by the evaluation sentiment dictionary that takes into consideration the dependency of the characteristic word and the probability belonging to the newspaper publisher of each characteristic word by the topic model and logistic regression. In the probability that each characteristic word belongs to a newspaper publisher, it is possible to quantitatively show the bias of a newspaper article that makes it easy to visually understand the relation of each newspaper publisher and each characteristic word by hierarchical clustering.
***Key Words** : Natural language processing, Topic model, Newspaper*

１．はじめに

インターネットや端末の普及につれて，ニュースサイト，ブログ，SNS 等の様々な形のメディアに触れる機会が増えている．総務省の平成 28 年版情報通信白書[1] では，インターネットの利用者数は年々増加し，インターネットメディアに対しての信頼度も増加している．

しかし，様々なメディアにおいて，執筆者の主観による情報の偏向性があり，その偏向性がどのような差異であるのかが一見して理解しづらいため，多種類のメディアから統合的に正確かつ公平な情報を受け取ることは難しい．

本研究では，比較的信頼度が高いニュースサイト(Web ニュース)に着目し，大手新聞社 4 社の朝日新聞，毎日新聞，産経新聞，読売新聞の Web ニュース記事を対象に，偏向性が解釈可能な”特徴語”に対する評価値を示すことにより，偏向性の可視化を行うことを目的とした．

２．関連研究

本研究に似た先行研究として，輪島らの研究[2]がある．これは潜在的ディリクレ配分法(LDA)[3]と評価表現辞書を用いて，質問(文書) の潜在的に存在するネガティブさ(深刻度)の傾向を明らかにするものである．この手法では，トピックに含まれる平均評価極性値を求め，その分布を俯瞰するものである．しかし輪島らの研究では，ネガティブ方向の極性値しか扱っておらず，また結果とした各トピックの平均評価極性値の分布は，偏向性を示すという目的には合わない．本研究とは，ポジティブな方向の極性も扱う他，偏向性が解釈可能な特徴語に対する評価値を結果とする点で異なる．

熊本らの研究[4]では，Web ニュース記事に対して受け手がどのように感じるかを推測するシステムを構築している．記事種毎に感情語による記事の感情尺度値等の評価値の比較という点では，本研究と似ているが，評価尺度は受け手に対する感情値であり，記事の客観性や正確性，公平性等の偏向性を俯瞰することを目的としている本研究と相反する．

市川らの研究[5]では，Web ニュースの報道の偏りを検知し可視化することを目的としており，この点は本研究と同じである．また，各新聞社の単語の扱われかたを示すという点も等しい．しかし，市川らの研究では，その手法として Word2Vec を用いた語の分散表現を主成分分析と t-SNE により，2 次元に圧縮し，図示するというものである．この研究の問題点は，各新聞社の語の扱われ方について，語の分散表現間の距離感，つまり語の意味の近さを掴

むことはできるが，一見して新聞社間の偏りを理解させられるとは言い難い．本研究では，各新聞社の語の扱われ方，各新聞社間の偏りを一見して理解させるため，偏向性が解釈可能な特徴語に対する評価値を示すシステムの構築を行う．

３． 方法

本研究では，記事の偏向性を解釈可能な特徴語に対して，各新聞社がどのように扱っているかを表す評価値を示すシステムの構築を行う．本研究ではこの評価値を，特徴語に対する各新聞社の極性評価値平均と，極性を考慮した特徴語を含む記事の各新聞社に属する確率の 2 種類の評価値を用いる．特徴語に対する各新聞社の極性評価値平均を算出するまでの流れを Fig. 1，極性を考慮した特徴語を含む記事の各新聞社に属する確率を算出するまでの流れを Fig. 2 に示す．

Fig. 1，Fig. 2 における各処理は以下の節で説明する．

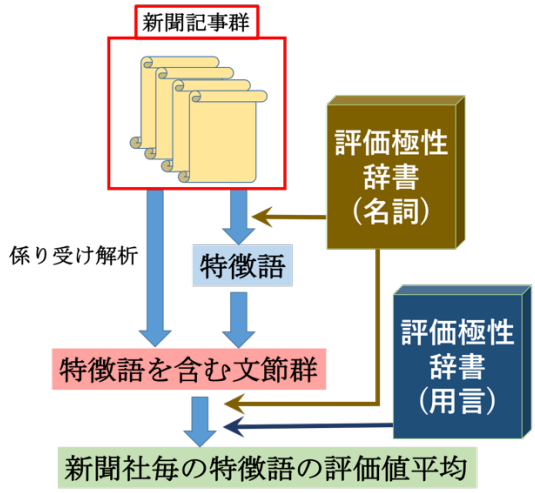

Fig. 1 特徴語に対する各新聞社の極性評価値平均を算出するまでの流れ

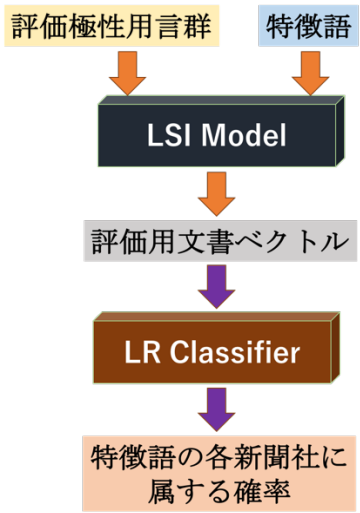

Fig. 2 極性を考慮した特徴語を含む記事の各新聞社に属する確率を算出するまでの流れ

（１）Web ニュース記事

各新聞社の Web ニュースの経済・政治・国際のカテゴリの記事を，2014 年の 7 月から 2015 年の 11 月までに取得したものから，ランダムに各新聞毎に 5610 件を抽出し，これを対象とした．

（２）形態素解析

対象となった記事群に対して，まず前処理として形態素解析を行い，トークン単位に分割を行った．形態素解析には，MeCab [6]を用い，新語に対応させるために mecab-ipadic-NEologd [7]を使用した．

（３）特徴語

本研究では記事の偏向性が解釈可能な語(特徴語)を評価極性を持ち，記事内で頻出し，出現する記事数が多い名詞と仮定する．

特徴語を決定するために，形態素解析を行った新聞記事d_jにおけるトークンt_iについて以下の式で与えられる$\mathrm{tfdf}_{i,j}$を定義する．

$$\mathrm{tfdf}_i = \mathrm{df}_i \sum_k \mathrm{tf}_{i,k}$$

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\mathrm{df}_i = \log \frac{|\{d: d \ni t_i\}|}{|D|}$$

$n_{i,j}$はトークンt_iの記事d_jにおける出現回数，$\sum_k n_{k,j}$は，記事d_j におけるすべてのトークンの出現回数の和，$|D|$は総記事数，$|\{d: d \ni t_i\}|$はトークンt_iを含む文書数である．つまりtfdf_iは記事内で頻出し，かつ出現する記事数が多いトークンが高い値を示す．

語が評価極性をもつかどうかの判断のために，極性評価辞書[8],[9] を用いた．この評価極性辞書に含まれ，かつtfdf_i値が高い名詞を特徴語とし，用言を評価極性用言群とした．

（４）係り受け解析

新聞社毎の特徴語の評価極性値を算出する前に，記事から係り受け解析を行い，特徴語の係る文節群を抜き出す．係り受け解析には CaboCha [10] [11] を使用した．これにより，特徴語のかかるトークン群(文節) を抽出した．この文節中のトークンの極性値(ポジティブの場合 1,ネガティブの場合-1 をとる)の合計が文節の極性評価値となり，新聞社毎にこの合計をとることによって，各新聞社の特徴語の評価極性値を算出した．

（５）LSI モデル

特徴語の各新聞社に属する確率を求める際に，トピックモデルを使用した．トピックモデルは文書の潜在的な意味解析を行う際に用いる手法である．本研究では，トピックモデルの一つとして Latent Semantic Indexing(LSI) モデルを用いた．LSI は文書に含まれる語 M 語と，文書 N

個による MN 行列を特異値分解により，次元削減を行う．この削減された次元による空間が単語ベクトルと文書ベクトルによる潜在意味的な空間を表しており，概念空間と呼び，概念空間の基底をトピックと呼ぶ．

本研究では，文書単位を記事として扱う．まず全記事に含まれる全トークンの内，2 回以上出現し，かつ助詞，助動詞でないトークンについて ID を振り分ける．これにより，記事毎に各トークンの出現回数によって構成される文書のベクトルが与えられる．文書ベクトルの次元を M，全記事数を N とした場合，この文書ベクトルを並べた，MN 行列の特異値分解を行った．特異値分解により得られた行列によって，潜在意味解析を行いたい記事(以下クエリ記事) を，概念空間に写像してやることで，クエリ記事の潜在意味を表すベクトルを得ることができる．

（６）ロジスティック回帰モデル

概念空間に写像されたクエリ文書ベクトルが，どの新聞社に属するかの確率を求めたい．その際本研究では，ロジスティック回帰モデルを使用した．モデルのフィッティングには L2 正則化項を用いた．

また，LSI モデルとロジスティック回帰モデルの性能を示す事前実験として，新聞記事の新聞社推定を行った．新聞社毎の記事 5610 件のうち 4488 件をモデルフィッティング用に，1122 件をテスト用に分けて推定を行った．推定の流れはモデルフィッティング用の文書により，LSI モデルを作成し，全記事を概念空間に写像する．概念空間に写像したモデルフィッティング用文書ベクトルにより，ロジスティック回帰モデルのパラメータのフィッティングを行う．その後，概念空間に写像したテスト用文書ベクトルがどの新聞社に属するかを決定した．この実験では，accuracy は 0.8077 という結果が得られた．なお LSI の概念空間の次元数(トピック数)は 1000 とした．本研究では，この実験でフィッティングした LSI モデルとロジスティック回帰モデルを使用する．

（７）特徴語の各新聞社に属する確率

特徴語を表すようなクエリ文書を作成し，LSI モデルとロジスティック回帰モデルによって各新聞社に属する確率を算出する．特徴語を表すようなクエリ文書は，ポジティブ，ネガティブの極性を考慮した文書とする．本研究ではこのクエリ文書を，特徴語と極性値が与えられている用言(評価極性用言群) を含んだ文書とした．評価極性用言群は，特徴語の抽出と同様に，用言の評価極性辞書を用いて 30 語抽出を行った．特徴語を表すようなクエリ文書として，極性毎の評価極性用言群と，特徴語を用言の語数 30 語分を含む文書とした．つまり，クエリ文書は極性毎，全特徴語 60 語毎に作成されるので，120 件のクエリ文書が作成される．これらのクエリ文書を LSI モデルとロジスティック回帰によって各新聞社に属する確率を算出する

４． 評価実験

特徴語に対する各新聞社の極性評価値を，極性語数で割ったもの，つまり 1 文節あたりの極性評価値を，ポジティブ極性を Table 1, ネガティブ極性を Table 2 に示す．

Table 1 ポジティブな特徴語の文節あたりの極性評価値

	asahi	mainichi	sankei	yomiuri
平和	0.947573	0.670360	0.418919	1.039387
株	-0.056680	-0.534714	-0.083851	0.032129
活動	0.280075	0.007177	0.012658	0.291837
景気	0.457055	0.175719	0.428894	0.520979
支持	0.127789	-0.058824	0.177719	-0.042079
協力	0.432348	0.159223	0.240000	0.249221
資金	0.124277	0.007843	0.260684	0.112727
金	0.250000	0.226833	0.442308	0.372340
技術	0.340278	0.424165	0.567358	0.523404
回復	0.573034	0.511070	0.428291	0.668555
安全	0.245678	0.223810	0.127297	0.350312
長	0.113229	-0.063353	0.072695	0.101025
認識	-0.176282	-0.054466	-0.003448	-0.007143
支援	0.145963	0.011299	0.188366	0.116972
最高	-0.084034	-0.137767	-0.191824	-0.018405
ポイント	0.006110	-0.020183	-0.141058	0.006780
改善	0.136951	0.287582	0.280872	0.222222
利益	0.267516	0.185268	0.123711	0.126506
情報	0.122924	0.002039	0.116162	0.023055
成長	0.179348	0.164122	0.227571	0.302721
実現	0.060847	0.204626	0.146186	0.156863
大統領	0.045793	-0.051910	-0.078669	-0.040183
力	0.136364	0.109974	0.228528	0.160156
期待	0.109005	0.078056	0.052314	0.013736
目標	0.257246	0.219731	0.267606	0.315018
規模	-0.086076	-0.139415	-0.096515	-0.139665
得	0.100529	0.054124	0.102190	0.055147
ため	0.075023	0.025424	0.058735	0.023315
事実	-0.141221	-0.142157	-0.148855	-0.172932
改正	0.080745	0.110887	0.101124	0.106977

Table 2 ネガティブな特徴語の文節あたりの極性評価値

	asahi	mainichi	sankei	yomiuri
事態	0.224265	0.034043	-0.297101	0.016667
露	-0.454545	-0.077403	-0.106667	0.006623
被害	0.068852	-0.182540	0.208134	0.068027
戦争	-0.006024	-0.140389	-0.078313	0.186147
テロ	-0.313472	-0.505598	-0.177489	-0.408497
容疑	-0.348039	-0.590857	-0.556818	-0.362637
戦闘	-0.220833	-0.236593	-0.138095	0.039841
殺害	-0.235000	-0.537634	-0.420455	-0.358268
過去	0.227477	0.009724	0.273273	0.119850
死亡	-0.309028	-0.509574	-0.285714	-0.448931
増税	-0.061856	0.118380	0.180723	0.089385
拘束	-0.032258	-0.228296	-0.022901	-0.121622
事件	-0.404851	-0.546381	-0.624031	-0.564688
警戒	0.105634	-0.094624	-0.052083	0.038462
過激	-0.489489	-0.428150	-0.296820	-0.457529
対立	-0.026042	-0.034091	0.040000	-0.148718
価格	-0.021531	-0.164038	-0.026711	-0.127572
下落	0.058559	-0.106383	0.000000	-0.009346
懸念	-0.013550	-0.169318	-0.075949	-0.078947
反対	0.046392	0.019833	0.121406	-0.008163
反発	-0.047244	-0.066556	-0.050000	0.050420
要求	-0.030973	-0.095915	0.026718	-0.023729
取引	0.019746	0.043689	0.093863	-0.013699
批判	-0.066079	-0.069573	0.016701	-0.040816
危機	-0.152482	-0.124481	-0.098684	-0.069767
事故	-0.391304	-0.464286	-0.395522	-0.410000
課題	0.179825	0.132992	0.207547	0.149780
問題	-0.021359	-0.058312	0.001458	-0.044719
訴え	0.010363	-0.025292	0.020747	-0.031674
指摘	-0.014881	-0.024021	-0.027816	-0.029557

特徴語の各新聞社に属する確率のヒートマップを，ポジティブ極性評価用言群で表したポジティブな特徴語によるものを Fig. 3，ポジティブ極性評価用言群で表したネガティブな特徴語によるものを Fig. 4，ネガティブ極性評価用言群で表したポジティブな特徴語によるものを Fig. 5，ネガティブ極性評価用言群で表したネガティブな特徴語によるものを Fig. 6 に示す．ヒートマップの周りの樹形図は，新聞社毎，特徴語毎に，ベクトルのユークリッド距離によって階層的クラスタリングを行ったものである．

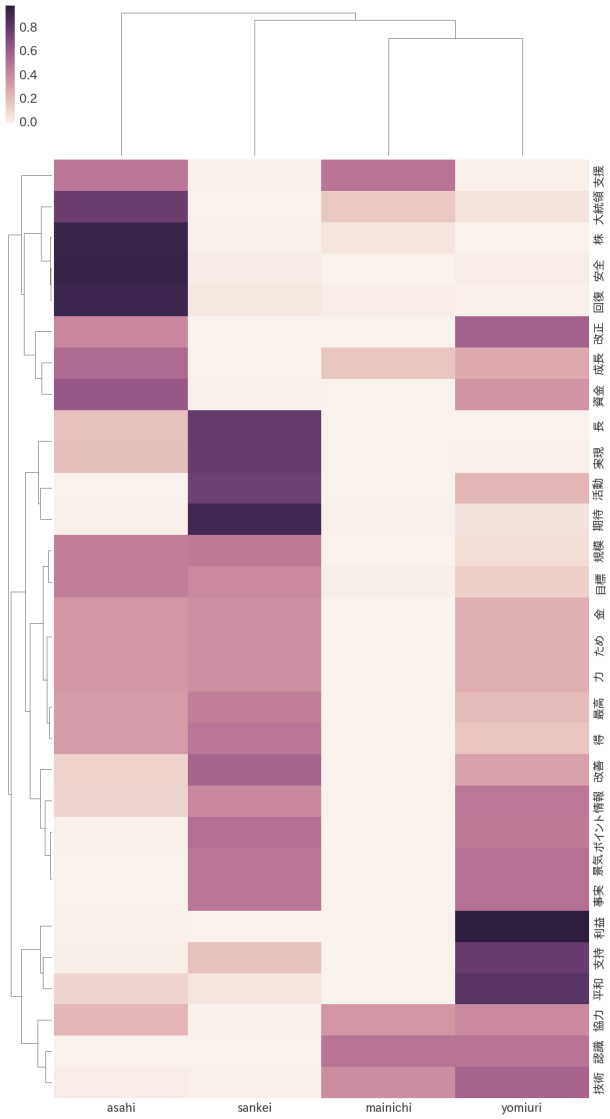

Fig. 3 ポジティブ極性評価用言群で表したポジティブな特徴語の各新聞社に属する確率

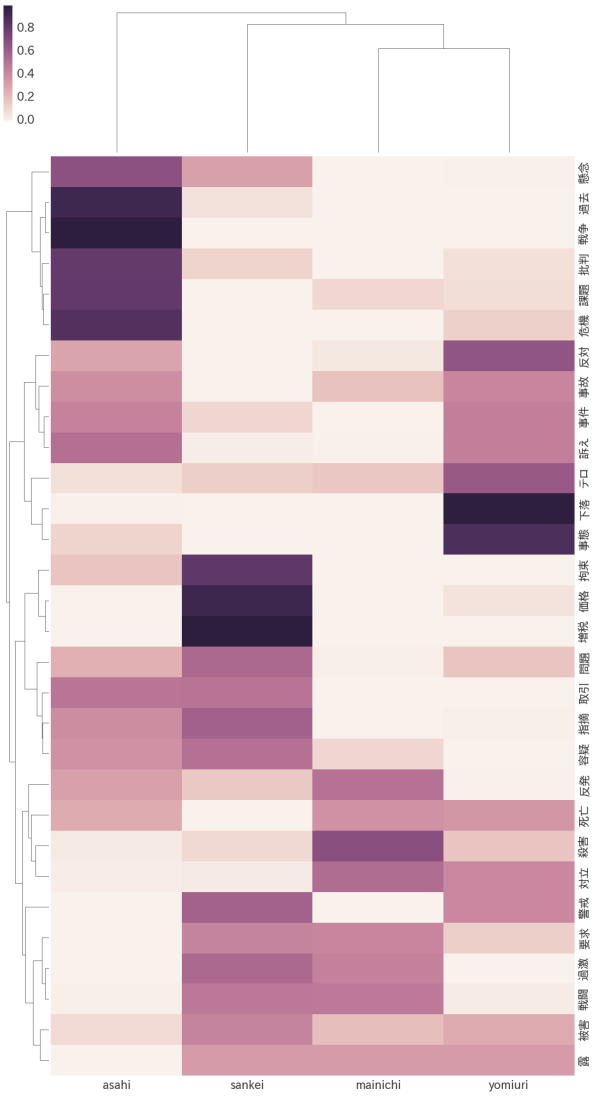

Fig. 4 ポジティブ極性評価用言群で表したネガティブな特徴語の各新聞社に属する確率

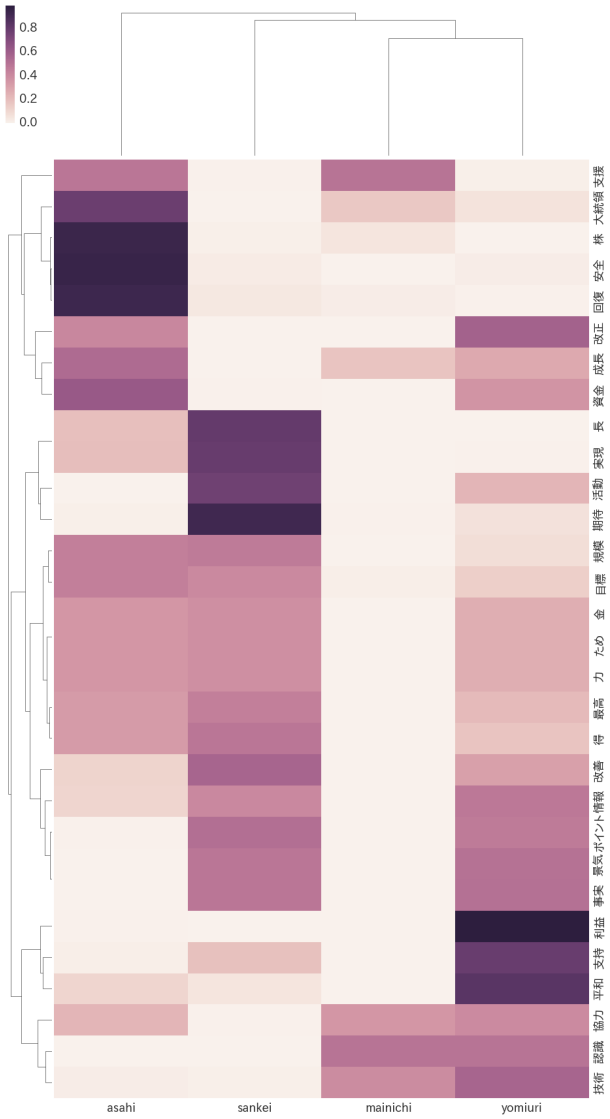

Fig. 5 ネガティブ極性評価用言群で表したポジティブな特徴語の各新聞社に属する確率

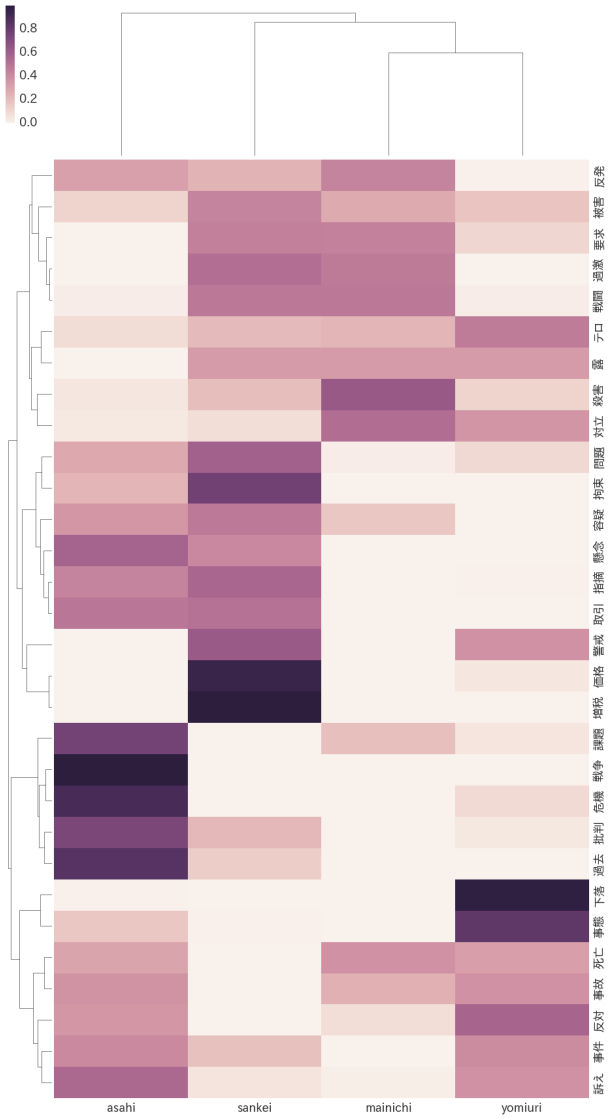

Fig. 6 ネガティブ極性評価用言群で表したネガティブな特徴語の各新聞社に属する確率

５． 考察

特徴語に対する新聞社毎の極性評価値は，それぞれの特徴語に対してどのような修飾語，述語等を用いて書いてあるかを表している．朝日新聞では"テロ"という単語が-1.762，"事件" という単語は-2.991 と"事件" の方がよりネガティブに書いているのに対して，毎日新聞では"テロ" が-5.448，"事件" が-3.347 と，"テロ" のほうがよりネガティブに書いている．他の例としては，"増税" という単語に対して，毎日新聞，産経新聞，読売新聞ではプラスの値を示すが，朝日新聞ではマイナスの値を示している．これは"増税" という単語に対して朝日新聞は，他の特徴語よりはポジティブな単語を使用していないことが分かる．特徴語の各新聞社に属する確率の結果を見てみると，例えば，"平和"，"支持"，"利益" という単語を含む記事はポジティブ，ネガティブの両極性について読売新聞に属する確率が高いが，ネガティブな記事であると"利益" と

いう単語が読売新聞に属する確率が他の 2 語より高く，ポジティブな記事であると，その差があまり無い．つまり，"利益" という単語について"支持" や"平和" よりもポジティブに書くより，ネガティブに書いたほうが，より読売新聞らしい記事であるということが言える．他の例としては，"殺害" という単語について，産経新聞と読売新聞で比較すると，ネガティブな用言を含む"殺害" の記事はどちらかというと産経新聞に属しやすく，ポジティブな用言を含む"殺害" の記事は読売新聞のほうが属しやすいと考察できる．新聞社毎の樹形図を見ると，全 60 語の特徴語について，基本的に毎日新聞と読売新聞の距離が一番近いが，ポジティブな用言を含むポジティブな特徴語では，毎日新聞と産経新聞のほうが近い．これは，基本的に毎日新聞と読売新聞が似ているが，ポジティブな話題に対してポジティブに書く記事では，毎日新聞と産経新聞のほうが似ていると言える．

６．結論

新聞社 4 社の Web ニュース記事を対象に，特徴語に対する新聞社毎の極性評価値と，特徴語の各新聞社に属する確率という 2 つの指標により，新聞社の偏向性の定量評価を行った．その結果，各特徴語に対する扱いが新聞社毎に異なっていることにより，偏向性が確認できた．また，各新聞社の相似性やその新聞社らしさ等も確認，考察ができる結果となった．しかし，特徴語の各新聞社に属する確率で，ポジティブとネガティブの評価極性用言による差が比較的少ないという問題点がある．この問題点を解決するためには，特徴語を表すクエリ文書の定義をより正確に定義する必要がある．他の問題点としては，特徴語が偏向性を評価するための基底として適切かどうかを検証していないため，その評価，検証を行うとともに，特徴語の抽出方法についても考察する余地があるように思われる．また，2 つの指標について偏向性が確認できるものの，一見しては理解しづらく，2 つの指標についての表示方法について改善すべきである．

謝辞：本研究にあたり，全般にわたるご指導をしてくださった彌冨仁准教授，および彌冨研究室の皆様に深く御礼申し上げます．

参考文献

1) 平成 28 年版情報通信白書. Technical report, 総務省, 2016.http://www.soumu.go.jp/johotsusintokei/whitepaper/h28.html.
2) 輪島幸司, 小河誠巳, 古川利博. 潜在的ディリクレ配分法を用いたネガティブ要因分析.第 6 回データ工学と情報マネージメントに関するフォーラム DEIM2014 講演論文集, pp.A9-3, 2014.
3) David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. Journalof Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
4) 輪島幸司, 小河誠巳, 古川利博. 潜在的ディリクレ配分法を用いたネガティブ要因分析.第 6 回データ工学と情報マネージメントに関するフォーラム DEIM2014 講演論文集, pp.A9-3, 2014.
5) 市川祐太. テキストマイニングを用いた新聞メディアの報道傾向検出への試み. Master's thesis, 法政大学大学院理工学・工学研究科, 2016. (未公刊).
6) 工藤拓，山本薫，松本裕治. Conditional Random Fields を用いた日本語形態素解析. 情報処理学会研究報告自然言語処理（NL）, Vol. 2004, No. 47, pp. 89{96, may 2004.
7) Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.
8) 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203-222, 2005.
9) 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会発表論文集, pp. 584-587, 2008.
10) 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. Vol. 43, No. 6, pp. 1834-1842, 2002.
11) Yuji Matsumoto Taku Kudo. Japanese dependency analysis using cascaded chunking. In CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops), pp. 63-69, 2002.